ISSN 2229-5518

A DKIM based Architecture for Combating Good Word Attack in Statistical Spam Filters

Kashefa Kowser.K, Saruladha.K, Packiavathy.M

Abstract— Abuse of E-Mail by unwanted users causes an exponential increase of E-Mails in user mailboxes which is known as Spam. It is an unsolicited commercial E-mail or unsolicited bulk E-Mail produces huge economic loss to large scale organizations due to high network bandwidth consumption and heavy mail server processing overload. Statistical spam filters could be used to categorize incoming E-Mails into legitimate and spam but they are vulnerable to Good Word attack which obfuscates "good words" in spam messages to make it legitimate. This paper attempts for a counterattack strategy to eradicate insertion of good words by proposing architecture of enhanced DKIM (DomainKeys Identified Mail) as a solution. Our experimental result shows that DKIM serves to be the best as it incorporates sender evidence with random values in the E-Mail messages which is critical for the spammers to evade E-Mail filtering process. The misclassification of the spam E-Mail as legitimate E-Mail would reduce the performance of text classifiers. As the number of E-Mail increases, the misclassification percentage decreases by using DKIM

Index Terms— spam filtering, good word attack, DomainKeys Identified Mail (DKIM)

1 INTRODUCTION

THE statistical spam filters use Machine Learning Techniques for automatically sorting text sets into categories from a predefined set. They are broadly classified into Reinforcement learning, supervised learning, semi-supervised learning and unsupervised learning. The learning method for each technique differs. In supervised learning method all training data are mostly labeled, unsupervised method train machines to learn by using unlabelled data, Semi-supervised learning technique uses both labeled and unlabeled data for training whereas reinforcement learning makes use of an agent to train data.

Text Categorization approach has considerable savings in labor power for organizing and handling text data than the knowledge engineering approach which requires data to be collected with the help of the domain experts either through direct interaction or through question raise with the help of the domain experts. Though Text Classification filtering Techniques is proven useful in statistical spam filters, spammers systematically modify the E-Mail messages and malicious contents enter the user's host bypassing the filters. One such type of attack is known as Good Word Attack in which spam messages are injected with enough good words which tends the text classifier system to classify a spam as a legitimate E-Mail. Spammers are explicitly trained to learn the features (keywords) which mostly occur in legitimate E-Mails and add those sets of good feature words(Most frequently occurring words in legitimate E-mails) to make the spam messages legitimate.

Also they append the spam keywords with spaces and punctuation symbols so that they are not filtered by the statistical spam filters. Even though a large body of research was proposed to this good word attack, there is paucity of misclassifications of features. DKIM [8] is a defense mechanism which uses digital signatures and guarantees authenticated E-Mail service. Further Domain Keys offers end-to-end integrity from a sender to the intended recipient with randomly generated evidence values.

This paper is organized as follows. Section 2 summarizes the related work, Section 3 discusses the architecture design of the proposed work, Section 4 discusses the experimental results and Section 5 is the conclusion.

2 RELATED WORKS

Enrico Blanzieri et.al presents an overview of machine learning applications for spam filtering and compares the different filtering methods. They also discuss other branches of anti-spam protection and use of various approaches in commercial and noncommercial anti-spam software solutions [1]

Fabrizo Sebastiani et.al compares the various automated approaches of text categorization algorithms in the way the classifiers are constructed and further evaluate the above said approaches for document indexing within the general machine learning Paradigm [2].

Sirisanyalak et. al uses an E-Mail feature extraction technique that extracts a set of four features and has used those features as input for spam detection model in artificial immune spam systems [3].

Gregory Wittel et. al examines the general attack method like common word attack and dictionary attack in

- *Kashefa Kowser.K is currently pursuing masters degree program in Information security in Pondicherry Engineering College, India, PH-9597652892. E-mail: kashefais@gmail.com*
- *Saruladha. K is currently working as a Assistant Professor and pursuing PhD in Pondicherry Engineering College, India, PH-9442396080. E-mail: charusanthaprasad@yahoo.com*
- *Packiavathy. M is currently pursuing masters degree program in Information security in Pondicherry Engineering College, India, PH-9943978904. E-mail: packiavathy_m@gmail.com*

the filter's features generation through tokenization or obfuscation along with the challenges faced by developers and spammers [4].

Daniel Lowd et. al describes the naïve bayes, maximum entropy statistical spam filters and evaluates the effectiveness of active and passive good word attacks on those filters [5].

Zach Jorgensen et. al applies multiple instance logistic regression on the multiple bags of instances (segments) and an E-Mail is classified as legitimate if all the instances in it are legitimate and as spam if at least one instance in the corresponding bag is spam [6]

Allman et.al [7] defines DKIM as a digital signature domain-level authentication framework that permits potential E-mail signers to publish E-Mail signing practices information for the E-Mail receivers to make additional assessments about messages using key server technology, public-key cryptography and Mail Transport Agents (MTAs) or Mail User Agents (MUAs).

Barry Leiba et.al focuses on verifying the digital signature that creates the evidence and ensuring both the sender and the recipient about the mail origin from where it says it does [8]

Erkut Sinan Ayla Havelsan et.al discusses intra-domain E-mail security system. It keeps E-Mail messages in corresponding mailboxes as encrypted messages. Trusted Mail Gateway process keeps encrypted E-Mail messages in mail boxes and records processing results in a database as notary information [9]

Ya-Jeng Lin et.al discusses the Lightweight, Pollution-Attack Resistant Multicast authentication scheme (PARM), which generates evidence that receivers can validate on a fast, per-packet basis. Fault-tolerance coding [10] algorithm which is discussed tolerates the loss of packet and signature amortization reduces the computation and communication overhead.

3 Proposed DKIM based Solution for GooD Word Attack

GoodWord attack is one of the most popular frequently employed attacks by spammers. The main issue in good word attack is that a spammer adds extra words or phrases to a spam message which are typically associated with legitimate E-Mail. Spam messages inflated with good words are more likely to bypass spam filters. Good word attack contains both passive and active attack. In passive attack, a word list is constructed by the attacker without any feedback from the spam filter. In active attack, text messages are sent to the filter to determine whether or not they are labeled as spam. So far, relatively little research has been done on how spam filters can be trained to account for such attacks. The misclassification of spam E-Mails as legitimate E-Mails (Good Word Attack) would reduce the performance of the text classifiers. This misclassification percentage could be reduced by the following methods.

- Frequent re-training of classifiers is an existing solution for combating good word attacks.
- Creating evidence of the sent E-mails so that the intercepted E-mails for injection of Good Words could be identified.

The first method though seen as a good solution the training of the text classifiers are to be done frequently and if the number of feature words in legitimate E-mails increases, the training time also increases.

This paper presents a novel approach for combating good word attack in statistical spam filters using DomainKeys Identified Mail (DKIM) based architecture. DKIM [8] defines a mechanism of cryptographically signing E-Mail messages permitting a signing domain to claim responsibility for the introduction of a message into an E-Mail system. Sender server publish public key in DNS (Domain Name Service) and then a sum using SHA256 [12] is calculated on selected header for sending an E-Mail. The sender generates a digital signature of the hashed message using RSA [11], a public key encryption scheme. The receiver server now looks public key up using DNS, decrypts the hash value and verifies the received sum. If the sum verifies, the sender server is verified, and the mail can be delivered. The proposed DKIM based architecture incorporates the sender evidence in the E-Mail messages to avoid the injection of good word thereby making the spam detection possible. Figure 1 shows the DKIM based architecture for combating good word attack. The following steps are to be followed for creation of evidence generation.

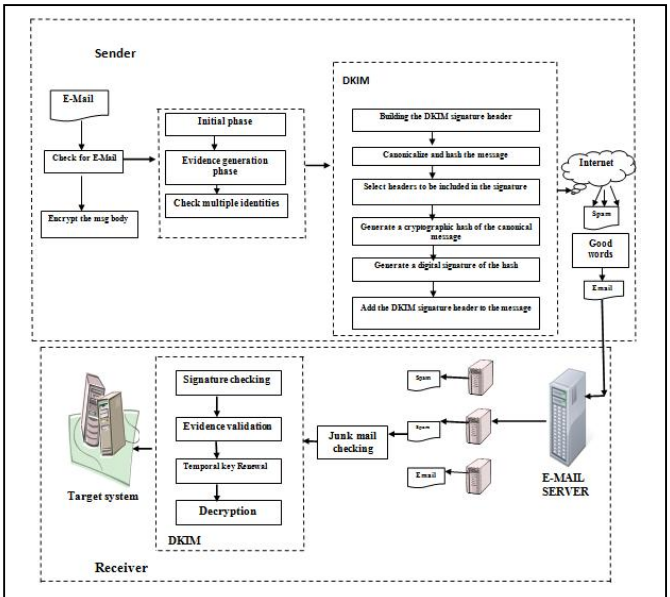

Figure 1: DKIM based architecture for combating good word

3.1 Evidence generation by the sender

In the first stage, Junk E-Mail is checked and controlled using its common spam characteristic features like discount, offer, bonus, money etc. The spam features can be identified with the help of message pre-processing used by a machine learning filter. Three main steps involved in the pre-processing of messages are

- Tokenization, decomposing the sentence into smaller units using punctuations, spaces etc., and extracting the features
- Lemmatization, reducing words to possible root word (e.g., "meeting" to "meet")
- Stop word removal, eliminating words like "to", "and", "for" that often occur in many messages. After data preprocessing if there is a spam keyword the DKIM header cannot be created otherwise the E-Mail is send for the encryption.

In the this stage, the body of the E-Mail is encrypted using Digital Signature Algorithm (DSA) [13] and a temporal key pair using one-way hash function is generated. Temporal key pair contains a Temporal Secret Key (TSK) chain and a Temporal Public Key (TPK). The sender creates the evidence of a packet from a TSK chain, and the receiver validates the evidence of a received packet with the TPK [10]. An attacker can convince receivers to accept a forged TPK if the sender not sign the TPK with the digital signature during distribution.

The sender generates the evidence which should be lightweight and fast for the receiver. The receiver should generate the evidence before sending a message to determine the validity of the packet and the sender needs to maintain a usage table for a given temporal key pair to find the number of times the column index of the TSK element array is used. The spammers may use multiple identities for good word attack in E-Mail. Hence, multiple identities of the user can be avoided by checking the user's address and time of sending the message.

Next stage is to construct the DKIM signature header in which the header will be covered by the signature. The domain and identity is to be signed in the header and the selector identifies the signing key. After constructing the DKIM signature header, the signer calculates a hash of the message body using hash algorithm. The next choice is to go into the signature header is the canonicalization algorithm for the headers and for the body. Canonicalization stage is used for minimal transformation of the message that will give its best chance of producing the same canonical value at the receiving end.

Header fields are the parts of the message that are most vulnerable to change in transit. In the next step, the signer can choose the header field to sign using DKIM by leaving insignificant header fields unsigned. This may increase the chances that the signature verified successfully.

SHA256, hash algorithm is used to generate a cryptographic hash for the canonical message. A hashing algorithm takes a variable length data message and creates a fixed size message digest. Then the signer signs the hash using the RSA encryption algorithm in the signature header, and adds it to the beginning of the message header fields. Finally the encrypted content will be added in the DKIM header. This completes the task of creation of evidence by the sender.

3.2 Evidence validation by the receiver

In checking signature phase, valid signature header must be checked by the receiver. The desired key identity is determined and retrieved from the specified key store. It is then validated and the public key is extracted from it. Policy retrieved from the receiver should be through the DNS query. From address has the domain for the query. In evidence validation phase, the receiver can use the TPK to immediately check the validity of the attached evidence when receiving a packet.

Algorithm at Sender Side
Input: DKIMBevidgen (E-Mail);
Output: Evidgen E-Mail

//Evidence attached E-Mail (Evidgen E-Mail)

1) Check the E-Mail.
2) Encrypt the message body using Digital Signature Algorithm (DSA).
3) Initialize Temporal Public Key (TPK) & Temporal Secret Key (TSK) for storing the evidence.
4) Create the evidence of a packet using TSK chain.
5) Calculating generation of evidence, hash & concatenates P with Q //P represents packet which is going to transfer, Q represents sequence number of the packet.
6) Append the evidence in the usage table //Usage table contain (TPK) and (TSK).
7) Generate n random numbers.
8) Hashing the random value using SHA256.
9) Building the DKIM header and canonicalize the message.
10) Selection of the header.
11) Digital signature of the hashed value using RSA.
12) Concatenate the message to DKIM header.

Algorithm at Receiver Side
Input: DKIMBevidcheck (Received E-Mail)
Output: Original E-Mail

13) Check the signature
14) Validate the evidence using TPK
15) Renewal of the used TSK elements.
16) Decryption of the encrypted message using Receiver's private key.

The attacker must generate proper evidence for a packet to forge, which is difficult without the knowledge of the TSK chain. The receiver must also maintain a usage table for each column index of the TSK elements array based on received packets like sender. Periodic renewal of used TSK elements ensures secure communications between the sender and its receiver. The final phase is to decrypt the encrypted message using the receiver's private key.

ISSN 2229-5518

4 EXPERIMENTAL SETUP

The implementation of the system is done in windows platform using JAVA on the publicly available spam corpus-Ling spam. The Ling spam corpus consists of 2171 legitimate E-Mail and 432 spam E-Mails in which 50% of the datasets is taken for implementation. The evidence generation value of random numbers contains E-Mail details, canonicalization value, the part which is generated by the header, selecting header in E-Mail details, SHA value, hash value and DKIM header value of the E-Mail in Table 1. Figure 2 represents the graph showing the misclassification percentage reduced by DKIM. As the E-Mail increases the misclassification percentage also increases. But the usage of DKIM decreases the misclassification percentage thereby combating the good word attack.

E-Mail Details	Header Generation	Canonicalization	Selecting header
To:dd@gmail.com From:ffff@rediff.com Subject:conference Date:Tue Apr 19 12:17:17 IST 2011 Message:International Conference on IS	i=@gmail.com	i=@gmail.com÷ c=simple/simple	i=@gmail.com÷ c=simple/simple t=Tue Apr 19 121717 IST 2011 h=ToFromSubjectDateMessage

Hash value	Sha value	DKIM header	Usage table
VG8= RnJvbQ == U3Via mVjdA == RGF0Z Q== TWVzc 2FnZQ ==	2047 06504745195334 42373552043352 37 13375151333937	i=@gmail.com§ c=simple/simple t=Tue Apr 19 121717 IST 2011 a=rsa-sha256 h=VG8=RnJvbQ==U3ViamVj dA==RGF0ZQ==TWVzc2FnZ Q== 2047 0650474519533442373552043 35237 133751513339370,×ä< "-&§Äü¯£šÅf¼),ÛUEř˜§ó™L ÷ãĩãóÊÌ Z™§	86:Q0ybWuUUZGu9kbUAMsp Xnv7I8vwtKrBLmWZfEGOD dY=

Table 1: DKIM based architecture for combating good word attack.

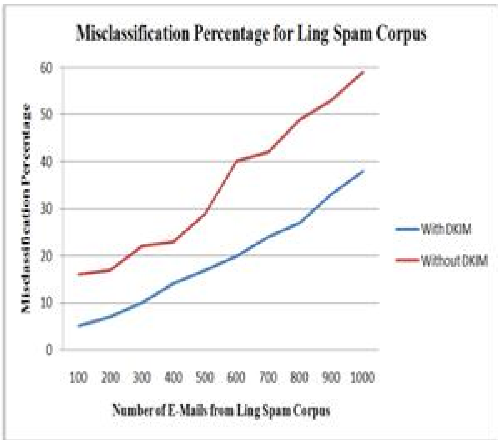

Figure 2: Misclassification Percentage for Ling Spam Corpus

5 CONCLUSION

This paper presented novel solution for good word attack by employing DKIM mechanism. It differs from the other counterattack strategy as it incorporates sender evidence in the E-Mail messages thereby making Spam detection possible. The result shows that the misclassification percentage decreases as the mail increases with the help of DKIM to eradicate the insertion of good words in spam E-Mail which makes as legitimate.

REFERENCES

[1] Enrico Blanzieri, Anton Bryl, "A Survey of Learning-Based Tec niques of Email Spam Filtering", January 11, 2008.

[2] Fabrizio Sebastiani, "Machine Learning in Automated Text Categorization", ACM computing surveys, Vol 34, No 1, 2002.

[3] B. Sirisanyalak and O. Sornil, "An artificial immunity-based spam detection system", Proc. of IEEE Congress on Evolutionary Computation, 2007, pp.3392-3398.

[4] Gregory Lee Wittel , "Evaluating and Attacking Statistical Spam Filtering Systems", Thesis, B.S. (University of California, Davis) 2002.

[5] D. Lowd and C. Meek, "Good word attacks on statistical spam filters", In Proceedings of the 2nd Conference on Email and Anti-Spam, 2005.

[6] Y. Zhou, Z. Jorgensen and M. Inge, "Combating Good Word Attackson Statistical Spam Filters with Multiple Instance Learning", Proc.of 19th IEEE International Conference on Tools with Artificial Intelligence, 2007, pp.298-305.

[7] Allman, E., Delany, M., and J. Fenton, "DKIM Sender Signing Practices", Internet Draft, http://www.ietf.org/internetdrafts/draft-allman-dkim-ssp-02.txt (work in progress), August, 2006.

[8] Barry Leiba, Jim Fenton, "DomainKeys Identified Mail (DKIM) Using Digital Signatures for Domain Verification", Journal of Foundations and Trends in Information Retreival, pp. 538-549, January 2008.

[9] Erkut Sinan Ayla Havesan, Ankara, Attila Ozgit, "An Architecture for End-to-End and Inter-Domain Trusted Mail Delivery Service", Communications of the ACM, pp. 24-33, February 2007.

[10] Ya-Jeng Lin, Shiuhpyng Shieh, Warren W. Lin, "Lightweight, Pollution-Attack Resistant Multicast Authentication Scheme", ASIAACCS'06, March 21-24, November 2006.

[11] http://www.di-mgt.com.au/rsa_alg.html

[12] http://www.ocean-logic.com/pub/OL_SHA256.pdf

[13] http://dsmc.eap.gr/members/pkitsos/papers/Kitsos_c09.pdf